Milestone 1: Index Construction Report

Information Analyst Option

This report presents the analytics for the inverted index constructed using the **Information Analyst** approach. This option is designed for groups with non-CS/non-SE students.

Approach Specifications:
• **Corpus:** Small portion of ICS web pages (analyst.zip) - ~2,000 pages
• **Index Storage:** Simple file-based storage (JSON format)
• **Memory:** Index fits entirely in memory during construction
• **Search Response Time:** Target < 2 seconds
• **Programming Level:** Introductory courses

Index Specifications:
• Tokens: All alphanumeric sequences
• Stop words: Not used (all words indexed)
• Stemming: Porter stemming algorithm
• Important words: Words in bold, headings (h1-h3), and titles are marked
• Term frequency: Calculated for each token in each document

Index Analytics

Metric	Value
Number of Indexed Documents	1,212
Number of Unique Tokens	13,126
Total Size of Index on Disk	22029.18 KB

Implementation Details

Index Structure:
The inverted index is stored as a dictionary in memory during construction, then saved to JSON format. Each token maps to a dictionary of document IDs, with each posting containing term frequency (tf) and an important flag.

File Organization:
• **inverted_index.json:** Contains the main inverted index structure
• **doc_mapping.json:** Contains mappings between URLs and document IDs

Processing Pipeline:
1. HTML content is parsed using BeautifulSoup to extract text and identify important elements
2. Text is tokenized into alphanumeric sequences
3. Tokens are stemmed using the Porter stemming algorithm
4. Term frequencies are calculated and stored in the inverted index (in memory)
5. Final index is saved to disk as JSON files

Memory Usage:
The entire index is held in memory during construction, which is feasible for the small corpus (~2,000 pages). The index size is approximately 14-22 MB, which fits comfortably in modern system memory.